6.1 什么是回归模型自变量的内生性？内生性会给参数的 OLS 估计带来什么影响？内生性产生的原因有哪些？

6.2 多元回归模型中的自变量有内生变量也有外生自变量，内生变量的内生性会对外生变量回归系数的 OLS 估计产生影响吗？

6.3 自变量的内生性为什么会引起回归系数 OLS 估计的不一致性？给出一个直观解释。

6.4 什么是工具变量估计法？为什么工具变量估计法能够解决回归系数 OLS 估计的不一致性？

6.5 设人体健康模型为

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + u$$

其中 *health* 表示健康指数，*age*、*weight*、*height* 和 *male* 分别表示年龄、体重、身高和性别虚拟变量（男性取值 1，女性取值 0），*work* 表示每周工作小时数，*exercise* 表示每周参加体育锻炼的小时数。

（1）*exercise* 与误差项相关吗？说出你的理由。

（2）设 *dhome* 和 *dwork* 分别表示家里和工作地点到最近的健身体育场馆的距离。*dhome* 和 *dwork* 能否作为 *exercise* 的工具变量？为什么？

（3）如果只允许用一个工具变量，选择哪个更合适？为什么？

6.6 设二元回归模型 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$ 中的解释变量均为外生变量。变量 Z 与 X_1 相关且与误差项 u 不相关。以 Z 为工具变量得出参数的工具变量估计是一致估计吗？工具变量估计与 OLS 估计那种方法更好？为什么？由此你会得到什么结论？

6.7 为了研究价格对香烟需求量的影响，建立香烟的需求模型。[①]

表 6.2

N	*D*	*P*	*Inc*	*Tax*	*Taxs*
1	101.1	158.4	83.90	40.5	41.9
2	111.0	175.5	46.00	55.5	63.9
…	…	…	…	…	…
47	115.6	166.5	32.61	41.0	50.4
48	112.2	158.5	10.29	36.0	36.0

表 6.2 给出了 1995 年美国 48 个州的香烟消费量、销售价格等有关数据。*D* 表示人均年香烟消费量（盒），*P* 为当年平均销售价格（$/盒），*Inc* 为人均年收入（$），*Tax* 表示销售税（对所有商品征收），*Taxs* 表示香烟销售税（只对香烟征收）。为了消除变量单位的影响，采用双对数模型。

（1）将香烟消费模型设定为

$$\ln(D) = \beta_0 + \beta_1 \ln(P) + u \qquad (6.14)$$

采用表中所给数据，用 OLS 方法估计出需求的价格弹性 $\hat{\beta}_{1OLS}^{(1)}$。这样设定的需求模型是错误设定的吗？为什么？会引起变量 $\ln(P)$ 的内生性吗？这样估计 β_1 合理吗？

（2）如果设定错误引起 $\ln(P)$ 的内生性，税收变量 *Tax* 能够作为 $\ln(P)$ 的工具变量吗？以 $\ln(Tax)$ 为工具变量计算价格弹性的工具变量估计 $\hat{\beta}_{1IV}^{(1)}$，与 $\hat{\beta}_{1OLS}^{(1)}$ 比较，并对比

[①] H. 斯托克、W. 沃森，《计量经济学》（第三版），沈根祥、孙燕 译，上海人民出版社，2012。

较结果进行解释。

（3）考虑到收入是影响需求的主要因素，将模型设为

$$\ln(D)=\beta_0+\beta_1\ln(P)+\beta_2\ln(Inc)+u \qquad (6.15)$$

计算（6.15）中的弹性系数 β_1 的 OLS 估计量 $\beta_{1OLS}^{(2)}$，与 $\hat{\beta}_{1OLS}^{(1)}$ 比较，并对比较结果进行解释。

（4）（6.15）中变量 $\ln(P)$ 是否仍然具有内生性？为什么？如果有，以 $\ln(Tax)$ 为工具变量估计出 β_1 的工具变量估计 $\hat{\beta}_{1IV}^{(2)}$，与 $\hat{\beta}_{1IV}^{(1)}$ 比较并对比较结果进行解释。如果只用一个工具变量，$\ln(Tax)$ 和 $\ln(Taxs)$ 哪个作为 $\ln(P)$ 的工具变量更为合适？

（5）以 $\ln(Tax)$ 和 $\ln(Taxs)$ 作为 $\ln(P)$ 的工具变量，用 TSLS 估计方法计算需求弹性的估计值 $\hat{\beta}_{1(2SLS)}$，与 $\hat{\beta}_{1IV}^{(2)}$ 进行比较并对比较结果进行解释。

（6）用基于回归的检验方法检验 $\ln(P)$ 的内生性。采用一个工具变量和两个工具变量分别进行检验。

6.8* 设一元线性回归模型为 $Y=\beta_0+\beta_1X+u$，X 具有内生性。设 Z 为 X 的工具变量，Z 为虚拟变量，取 0 和 1 两个值。

（1）证明：以 Z 为工具变量，β_1 的工具变量估计为

$$\hat{\beta}_{1IV}=(\overline{Y}^{(1)}-\overline{Y}^{(0)})/(\overline{X}^{(1)}-\overline{X}^{(0)})$$

其中 $\overline{Y}^{(0)}$ 与 $\overline{X}^{(0)}$ 分别表示对应 $Z=0$ 的 Y 和 X 的均值，$\overline{Y}^{(1)}$ 与 $\overline{X}^{(1)}$ 分别表示对应 $Z=1$ 的 Y 和 X 均值。对估计结果给出直观解释。

（2）如果 X 也是只取 0、1 值的虚拟变量，请结合一个实际例子对估计结果进行直观解释。

6.9* 在二元回归模型 $Y=\beta_0+\beta_1X_1+\beta_2X_2+u$ 中，X_1 为内生变量，X_2 为外生变量。变量 Z_1 和 Z_2 为 X_1 的工具变量。TSLS 估计法第一步将 X_1 对 Z_1 和 Z_2 和 X_2 进行回归，以回归拟合值 $\hat{X}_1$ 作为 X_1 的工具变量。为什么第一步回归中要包含 X_2？

◆ 参考答案

1. 内生性是指回归自变量和模型误差项相关。内生性会导致回归系数的 OLS 估计不一致性。内生性产生的原因有丢失相关变量、测量误差和联立方程，其中以丢失变量导致的内生性最为常见。
2. 如果外生自变量与内生自变量存在相关性，内生变量的内生性会引起外生自变量回归系数 OLS 估计的不一致性。渐近偏误的大小与自变量之间的相关程度有关，相关性越强偏误越大。
3. 由于和误差项相关，内生自变量变化将引起模型误差项的变化，进而间接引起因变量变会。回归系数衡量的是其他因素不变时自变量变化引起的因变量变化。对内生自变量来讲，不能保证其他因素不变。由于没能控制其他因素不变，回归系数的 OLS 估计，不仅包含内生变量变化引起的因变量变化，也包含内生自变量变化引起误差项变化带来的因 变量变化，不是回归系数的一致估计。
4. 工具变量估计法是借助于与内生自变量相关的另一个外生变量（工具变量）对回归系数进行估计的方法。工具变量变化引起内生自变量变化，从而引起因变量变化。由于与误差项相关，工具变量变化不会引起误差项变化。借助工具变量，能够阻断内生自变量变

化引起的误差项变化，正确估计出自变量对因变量的影响，得出回归系数的一致估计。

5. （1）相关。饮食（*diet*）是影响健康的重要因素，并且与锻炼相关。模型中没有包含 *diet*，相当于将其放入误差项，引起 *exercise* 与误差项相关。

（2）*dhome* 和 *dwork* 都可以作为 *exersice* 的工具变量，它们与饮食不相关，因此与误差不相关，为外生变量。同时与 *exersice* 高度相关：距离越近，锻炼越方便，锻炼越多。

（3）采用 *dhome* 作为 *exersice* 的工具变量更合适。工作期间不能参加锻炼，下班后时间有限。双休日和假期呆在家里，有更多锻炼时间，会选择离家近的场馆锻炼。

6. 不管自变量是否内生，工具变量方法得出的估计都是一致估计。当内生变量外生时，OLS 估计也是回归系数的一致估计，并且比工具变量估计的方差更小，更准确。因此，如果能够确定模型自变量不存在内生性，OLS 方法是估计模型的最优选择。

7. （1）在经济学中，需求和供给是相互影响的。模型自变量中没有包含的供给变量放入误差项 u 中。价格和供给的相关关系导致模型自变量 $\ln(P)$ 和误差项相关而成为内生变量，采用 OLS 估计出的 β_1 不是一致估计。

（2）销售税与供应无关，$\ln(Tax)$ 与误差项不相关，是外生变量。销售税与产品售价呈正相关：销售税越高，售价越高，因此 $\ln(Tax)$ 与 $\ln(P)$ 相关。$\ln(Tax)$ 可以作为 $\ln(P)$ 的工具变量。

（4）加入收入变量并没有反映出供给对需求的影响，（6.15）中的 $\ln(P)$ 仍然是内生变量。如果只用一个工具变量，$\ln(Taxs)$ 更为合适。作为专门对烟草征收的销售税，$\ln(Taxs)$ 与 $\ln(P)$ 的相关性更强。

（6）采用一个工具变量和两个工具变量进行内生性检验的区别，在于将内生变量 $\ln(P)$ 对所有外生变量回归以求得回归残差 $\hat{v}$ 时，回归自变量的多少。

8.（1）设 $Z=0$ 的样本个数为 n_0，$Z=1$ 的样本个数为 n_1，$n=n_0+n_1$ 为样本个数。由于 $\sum_i(Z_i-\bar{Z})(Y_i-\bar{Y})=\sum_i Z_i(Y_i-\bar{Y})$，$\sum_i(Z_i-\bar{Z})(X_i-\bar{X})=\sum_i Z_i(X_i-\bar{X})$，从（6.3）得出 β_1 的工具变量估计为

$$\hat{\beta}_{1IV}=\frac{\sum_i(Z_i-\bar{Z})(Y_i-\bar{Y})}{\sum_i(Z_i-\bar{Z})(X_i-\bar{X})}=\frac{\sum_i Z_i(Y_i-\bar{Y})}{\sum_i Z_i(X_i-\bar{X})}=\frac{\sum_{Z_i=1}(Y_i-\bar{Y})}{\sum_{Z_i=1}(X_i-\bar{X})}=\frac{\bar{Y}^{(1)}-\bar{Y}}{\bar{X}^{(1)}-\bar{X}}$$

。由 $\bar{Y}=(n_0\bar{Y}^{(0)}+n_1\bar{Y}^{(1)})n$ 得出

$\bar{Y}^{(1)}-\bar{Y}=\bar{Y}^{(1)}-(n_0\bar{Y}^{(0)}+n_1\bar{Y}^{(1)})n=(\bar{Y}^{(1)}-\bar{Y}^{(0)})n_0/n$

类似得出 $\bar{X}^{(1)}-\bar{X}=(\bar{X}^{(1)}-\bar{X}^{(0)})n_0/n$。带入 $\hat{\beta}_{1IV}$ 表达式得出结果。

（2）例如一项针对农民工的工作技能培训项目。$Z_i=1$ 表明第 i 个样本（农民工）有必要参加培训项目，$Z_i=0$ 表示没有必要参加（已经具有技能）。$X_i=1$ 表示参加项目培训，$X_i=0$ 表示没有参加项目培训。Y_i 表示农民工的年工资收入。此时 $\bar{X}^{(1)}$ 表示有必要参加培训的农民工中参加培训项目的比例，$\bar{X}^{(0)}$ 表示没有必要参加培训而参加培训项目的比例。$\bar{Y}^{(1)}$ 和 $\bar{Y}^{(0)}$ 分别表示有必要参加培训而参加了培训项目的农民工平均年工资和没必要参加而参加了培训项目的农民工平均年工资。回归系数 β 则表明了培训项目的效果（对农民工工资的影响）

9. 如果第一阶段的回归中不包含外生变量 X_2，对应的回归为 $X_1=\alpha_0+\alpha_1Z_1+\alpha_2Z_2+v$。$Z_1,Z_2$ 与误差项 v 不相关。得出的模型拟合值为 $\hat{X}_1=\hat{\alpha}_0+\hat{\alpha}_1Z_1+\hat{\alpha}_2Z_2$，拟合残差 $\hat{v}=X_1-\hat{X}_1$ 与 Z_1,Z_2 正交。由于自变量不包含 X_2，$\hat{v}$ 与 X_2 不正交。用 $X_1=\hat{v}+\hat{X}_1$ 带

入原模型得出第二阶段回归模型

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \\ &= \beta_0 + \beta_1 \hat{X}_1 + \beta_2 X_2 + u + \beta_1 \hat{v} \end{aligned}$$

模型的误差项为 $\beta_1\hat{v}+u$，由于 $\hat{v}$ 与 X_2 不正交，X_2 与 $\beta_1\hat{v}+u$ 相关，导致第二阶段回归模型存在内生自变量，OLS 估计的不一致性。如果第一阶段回归包含了 X_2，则 $\hat{v}$ 与 X_2 不相关，不会引起第二阶段回归的内生性。